15.4 A Low-Power Integrated x86-64 and Graphics Processor for Mobile Computing Devices

Srinivasa Rao Gutta[1], Denis Foley[2], Ajay Naini[1], Robert Wasmuth[3], Don Cherepacha[4]

[1]AMD, Hyderabad, India,
[2]AMD, Boxborough, MA,
[3]AMD, Austin, TX,
[4]AMD, Markham, Canada

AMD's first Fusion Accelerated Processor Unit (APU) codenamed "Zacate" (Fig. 15.4.1) combines a pair of x86 CPUs cores codenamed "Bobcat", 1MB L2 Cache, Client Northbridge (CNB), with a DirectX® 11 Radeon™ HD5000 graphics/multimedia controller on a single die. The CNB provides an interface to a single 64b DDR3 memory channel, which can operate at up to DDR3-1066. The Fusion architecture implements an efficient form of unified memory architecture (UMA) where a portion of system memory is reserved as graphics frame buffer memory. The graphics memory controller (GMC) arbitrates between graphics, video and display memory accesses and presents a well-ordered stream of system memory requests through the CNB over dedicated 256b wide read and write busses. These GMC requests bypass all of the CNB coherency mechanisms allowing for fast direct access to memory and exposing most of the available memory bandwidth (8.53GB/s). Compared to two chip solutions, use of the on-die integrated GPU significantly reduces memory latency, improves request ordering, and reduces power. The APU supports display formats including VGA, LVDS, Display Port, DVI or HDMI™. A ×4 Gen2 PCIe® Unified Media Interface (UMI) to an external Fusion Controller Hub (FCH) is supported for system I/O. An additional 4× PCIe Gen2 link supports I/O to external Discrete Graphics chip.

The SoC including the Bobcat cores is entirely synthesized from RTL and implemented using standard ASIC style SAPR flows. The die shown in Fig. 15.4.7 has over 450 million transistors and is manufactured using 10-layer metal 40nm bulk CMOS technology. As shown in Fig. 15.4.2, logic comprises 68% of die transistors, with approximately 56% HVT and 42% SVT. To limit leakage power, less than 2% LVT devices are used only in critical paths. The Bobcat core is implemented as a single physical tile consisting of 1.1 million instances and uses only 7 custom memory macros for branch predictor arrays, TLB, L1 tag and data arrays. The GPU core logical RTL is physically partitioned into multiple physical tiles with varying numbers of cell instances and memory macros generated using standard memory compilers. The GPU has zero grout space between the tiles and uses feed-through repeater bundles to connect the tiles.

The SAPR methodologies are carefully chosen at every step to have an optimal balance between performance, power and area. The CPU and GPU core operate on separate voltage supplies and support dynamic voltage and frequency scaling to optimize power consumption. The design is optimized for power and performance at discrete process and voltage points. The CPU is optimized for power and performance at 1.2V and 0.8V. The GPU is optimized at 3 voltage points 1.0V, 0.9V and 0.8V.

The design has 16 functional, 10 scan and 11 debug mode clocks and clock relationships which made the clock planning and implementation for the complete chip challenging. A Digital Frequency Synthesizer (DFS) is used to generate 9 functional clocks used by the CPUs, CNB, and the GPU as shown in Fig. 15.4.3. The system PLL provides 4 phase-offset references to the DFS which combines the phases to generate the required clock frequencies. The DFS allows the CPU to continue normal instruction execution during frequency transitions. The high speed CPU clock is distributed as a mesh network with 6-8 levels of local buffers and GPU clock (SCLK) is implemented with X-tree & H-tree topologies feeding into the Clock Tree Synthesis (CTS) branches inside a physical tile. All other clocks are implemented as balanced buffer trees from DFS to the physical tile. The CNB and GPU clocks are balanced globally with respect to each other within a ½ cycle of the fastest clock using programmable delay buffers pre-placed beside the global trees.

On-die power gating is implemented using a leakage optimized HVT PFET header switch to turn off the VDD rails. The header uses parallel stacked PFET transistors with separate enables for each PFET as shown in Fig. 15.4.4. The enables are daisy chained and have programmable delay to control the timing. This helps limit the inrush current and voltage transients. In contrast to other x86 designs using power-gating ring [1,2], the global power grid in the GPU uses M7 & M8 with headers placed on a regular grid as shown in Fig. 15.4.4. The distance between the headers is calculated as a function of minimum horizontal M7 grid pitch Y. The Memory macro power gating is implemented using headers arranged as a ring in 1 or 2 rows around the macro. The grid is designed to drive a current density of 0.5A/mm^2 and achieve a 2% static IR drop. The total gate width of GPU headers is 1.93m and resistance is 0.6mΩ. The CPU power grid is designed to drive higher current density and minimize frequency impact due to voltage drop associated with the header.

Bobcat and its L2 cache together share a power island as shown in Fig. 15.4.5. Both core power islands are supplied by a single variable VDD rail. Core clocks can be varied independently and VDD voltage is selected to be the lowest voltage required to support the highest of the selected core clocks. If either of the cores is idle its caches can be flushed, its state saved to memory and power to the core island gated off. If both cores are idle and power gated, VDD can be further reduced to eliminate leakage in even the power gating structures. GPU and CNB share a variable VDDNB rail. GPU and CNB clocks are varied independently based on activity. VDDNB is selected based on the highest voltage request from either GPU or CNB. The GPU supports several power islands allowing for driver controlled power gating of the video acceleration engine, and independent dynamic power gating of the GFX core and the GMC. The display controller supports a static screen refresh stutter mode where the controller requests data periodically from memory. Between requests the memory is kept in self-refresh and the CNB enters a low power state. By stuttering the display requests to memory, the time spent in self-refresh for the memory can be maximized, and the time that clocks are running in the CNB can be minimized. The combination of power savings techniques yields an APU with an average MobileMark® (MM07) power consumption of less than 1.8W.

As shown in Fig. 15.4.6, the Bobcat core provides nearly 90% of mainstream CPU performance [3] with approximately 50% reduction in area and dynamic power compared to the same reference core (excluding L2) if scaled to the same technology. The front-end contains a 32k 2-way L1 instruction cache with a return stack, indirect dynamic and advanced conditional branch predictors with a high-capacity ITLB. The pipeline supports out-of-order execution, retiring up to two complex uOPs per cycle – a significant differentiating factor compared with simpler in-order CPUs. The load-store unit enables out-of-order accesses to an 8-way 32KB data cache with a high-capacity two-level DTLB, and 8-stream data pre-fetcher. The bus unit allows for 8 outstanding data accesses and 2-outstanding fetch accesses with a 512KB 16-way L2 cache with optimized data buffering to the CNB for high memory bandwidth.

The GPU delivers a peak of 80GFLOPS, which provides sufficient compute capability for gaming or compute applications enabled by DirectCompute and OpenCL™. As shown in Fig. 15.4.6, the power efficient APU outperforms current mainstream UMA chipsets. The low-latency, high bandwidth GMC memory interface increases performance per Watt by allowing a higher percentage of the power to be used for computation rather than in moving large amounts of data across high power chip-to-chip interconnects. The advanced, multi-level memory arbitration units resolve performance issues with previous Integrated Graphic Processors (IGP) by enabling the GPU to achieve memory efficiency similar to a discrete GPU without compromising CPU performance. GPU memory accesses are sent in efficient streams of reads and writes to avoid DDR penalties while CPU accesses are optimized for low read latency. These improvements yield a 17% increase in Zacate UMA memory efficiency compared to its two-chip predecessor.

Acknowledgments:
The authors would like to acknowledge the talented AMD Engineering community for their contribution to the SoC.

References:
[1] R. Jotwani, et al. "An x86-64 Core Implemented in 32nm SOI CMOS", *ISSCC Dig. Tech. Papers,* pp.106-107, Feb. 2010
[2] T. Fischer, et al. "Design Solutions for the Bulldozer 32nm SOI 2-core processor module in an 8-core CPU". Accepted to ISSCC 2011.
[3] M. Golden, et al., "A 2.6GHz Dual-Core 64bx86 Microprocessor with DDR2 Memory Support," *ISSCC Dig. Tech. Papers*, pp.325-332, Feb. 2006.

 978-1-61284-302-5/11/$26.00 ©2011 IEEE

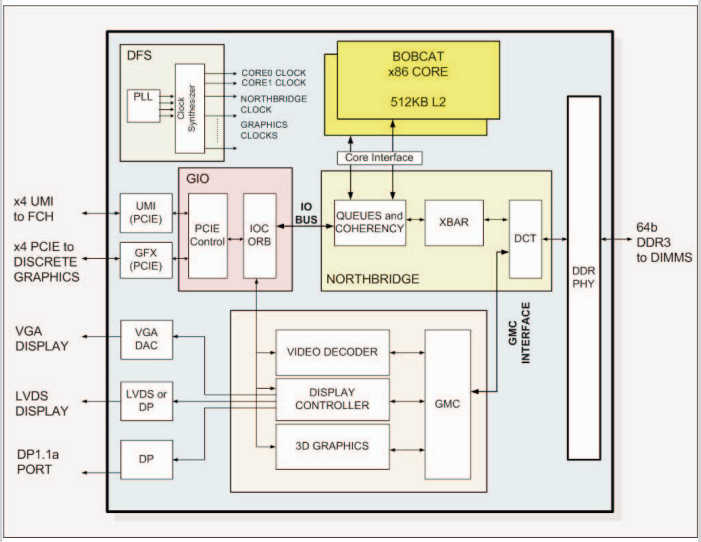


Figure 15.4.1: "Zacate" Architectural Block Diagram.

Instance Count	
Macro Instances	225
Memory Instances	866
Sequential Instances	2473912
Logic Instances	8080771
Inv/Buf instances	3089581
Total Instance count	13645355
Memory Bits	
Total Memory Bits in GPU	10559376
Total Memory Bits in Core+L2	12230872
Total Memory Bits in SOC	22790248
Transistor Count	
Memory Transistors	144200024
Standard cells (HVT, LVT, SVT)	305330039
IO Transistors	1828283
Total Transistors in SOC	451358346

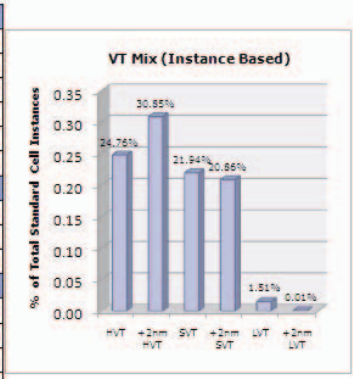


Figure 15.4.2: "Zacate" SoC Statistics.

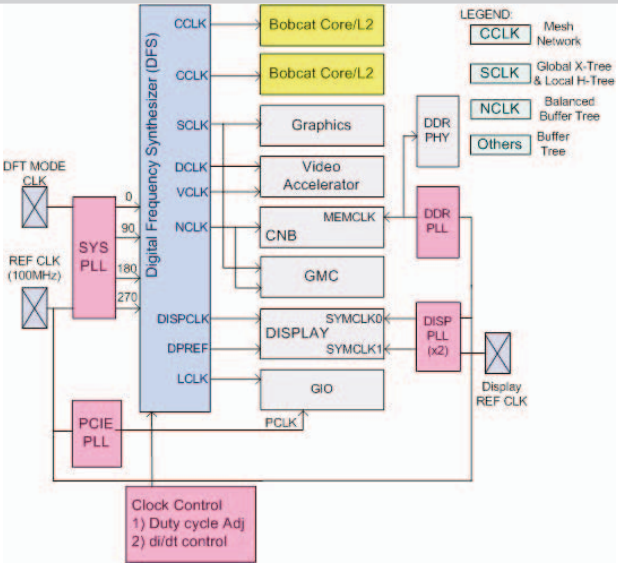


Figure 15.4.3: Clock Generation.

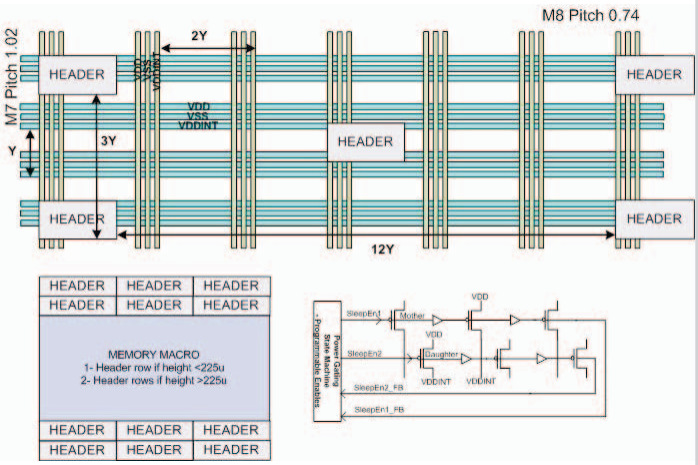


Figure 15.4.4: Power Gating Methodology.

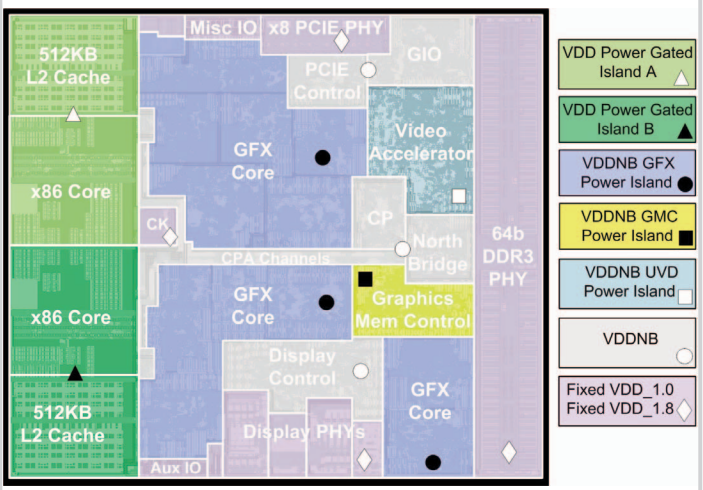
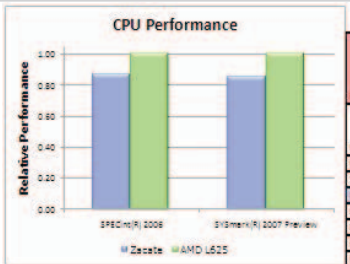
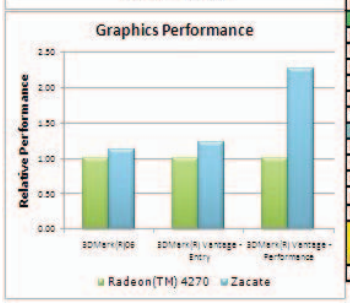


Figure 15.4.5: "Zacate" Die picture highlighting Power Islands and Logical Units.





System Configuration

Platform	AMD Inagua Reference Platform [Zacate]	AMD Guam Reference Platform [Radeon(TM) 4270]
CPU/APU	Zacate Engineering Sample	X920 Engineering Sample
Northbridge	NA	RS880
Southbridge	SB800	SB810
CPU Core Details		
CPU Speed	1.6 GHz	2.3 GHz
CPU Cores	2	4
HT Speed	NA	1800
HT Width	NA	16
System Memory Details		
Memory Speed	DDR3-1066	DDR3-1333
Memory Size per DIMM	2048 MB	2048 MB
Num Channels	1	2
DIMMs per Channel	2	1
Ranks per DIMM	2	2
Total Memory Size	4096 MB	4096 MB
Graphics Core Details		
Frame Buffer Size	384 MB	256 MB
Graphics Mode	Integrated	Integrated
Graphics Engine Clock	500 MHz	590 MHz
Local Memory Clock	NA	400 (64 MB SP)
Driver Version	8.79-1006	8.73 (5/13/2010)
OS Details	Windows(R) 7 (64-bit)	Windows(R) 7 (32-bit)
Aero	Enabled	Enabled

Figure 15.4.6: "Zacate" CPU & Graphics Relative Performance.

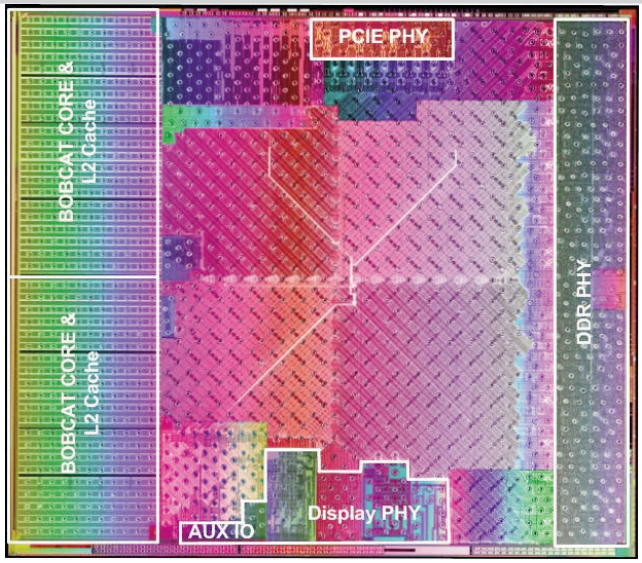


Figure 15.4.7: "Zacate" Die Photograph.

978-1-61284-302-5/11/$26.00 ©2011 IEEE